# Bleed-Through Removal in Document Images

**A. Yuvaraj**

**Raju Balan**


**Department of Computer Science and Engineering**
**National Institute of Technology Rourkela**
**Rourkela – 769 008, India**

# Bleed-Through Removal in Document Images

*Thesis submitted in partial fulfillment*
*of the requirements for the degree of*

## Bachelor of Technology

*in*

## Computer Science and Engineering

*by*


**A. Yuvaraj**
**(Roll: 10606027)**

**Raju Balan**
**(Roll: 10606058)**




**Department of Computer Science and Engineering**
**National Institute of Technology Rourkela**
**Rourkela – 769 008, India**


**May 2010**

Department of Computer Science and Engineering
**National Institute of Technology Rourkela**
Rourkela-769 008, India. www.nitrkl.ac.in

**Pankaj Kumar Sa**
Assistant Professor

May 07, 2010

# Certificate

This is to certify that the work in the thesis entitled *Bleed-Through Removal in Document Images* by *A. Yuvaraj & Raju Balan*, is a record of an original research work carried out by them under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering at the National Institute of Technology, Rourkela. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

***Pankaj Kumar Sa***

# Acknowledgment

This thesis has benefited in various ways from several people. Whilst it would be simple to name them all, it would not be easy to thank them enough.

We would like to gratefully acknowledge the enthusiastic supervision and assistance of *Prof. Pankaj Kumar Sa* throughout this work. His constant support and unstinting guidance has been always been an immense source of motivation and encouragement.

We are very much indebted to *Prof. Banshidhar Majhi*, Head-CSE and *Prof. P. M. Khilar* for allotting us this project and also for the resources and facilities that were made available to us whenever we needed the same.

Thanks are due to *Mr. Hemant Mishra*, *Mr. M. Madhan* and other staff of the Biju Pattnaik Library of NIT, Rourkela for graciously providing us with all facilities and access to the resources vital to the completion of this project.

Our sincere thanks to *Prof. S.K. Rath*, *Prof. S. K. Jena*, *Prof. B. D. Sahoo*, *Prof. A. K. Turuk*, *Prof. D. P. Mohapatra*, *Prof. S. Chinara*, *Prof. R. Baliarsingh*, *Prof. K. S. Babu*, *Prof. S. Mohanty* and *Prof. S. K. Panigrahi* for being our knowledge resource. Their help can never be penned in words.

We would like to thank all our friends for helping us and would also like to thank all those who have directly or indirectly contributed in the success of our work.

Last but not the least, big thanks to NIT Rourkela for providing us such a platform where learning has known no boundaries.

*A. Yuvaraj*
*Raju Balan*

# Abstract

When documents are written on both sides, quite often ink bleeds through the paper. This is a common phenomenon with old documents and low quality paper. With the presence of increased bleed-through, reading and deciphering the text becomes tedious. This thesis implements algorithms for reducing bleed-through distortion using techniques in digital image processing. A comparative study of three methods has been performed with the first being basic enhancement through thresholding. The next two methods are founded on a registration process which aims at working on both the recto and verso sides simultaneously. Firstly, both sides of the documents are digitized. The verso is then flipped and corrected so as to correspond to the coordinates of one side exactly with the coordinates of the original writing on the other which is done using an affine transformation of six parameters. The parameters are found by optimizing the alignment process. A restoration algorithm is then applied to remove bleed though areas on the desired side page. The third method proposes the use of cross correlation to handle the registration process. The accuracy of bleed-through correction is found to be largely dependent on the accuracy of alignment of the documents.

# Contents

# List of Figures

# Chapter 1

# Introduction

## 1.1 Image Processing

The sense of vision has been one of the most vital senses for human survival and evolution. A human uses the visual system to see or acquire visual information, perceive, i .e. process and understand it and then deduce inferences from the perceived information. Image processing deals with automating the process of gathering and processing visual information. The process of receiving and analysing visual information by a digital computer is called *digital image processing.*

An image may be described as a two-dimensional function $I$.

$$I = f(x, y) \tag{1.1}$$

where $x$ and $y$ are spatial coordinates. Amplitude of $f$ at any pair of coordinates *(x, y)* is called the intensity $I$ or gray value of the image. When spatial coordinates and amplitude values are all finite, discrete quantities, the image is called a digital image. Digital image processing may be classified into various sub-categories based on methods whose: [1]

- input and output are images
- inputs may be images where as outputs are attributes extracted from those images.

Following is the list of different image processing functions based on the above two classes.

- Image Acquisition
- Image Enhancement
- Image Restoration
- Color Image Processing
- Multi-resolution Processing
- Compression
- Morphological Processing
- Segmentation
- Representation and Description
- Object Recognition

For the first seven categories the inputs and outputs are images where as for the other three the outputs are attributes of the input images. With the exception of image acquisition and display, most image processing functions are implemented in software. Image processing is characterized by specific solutions, hence the technique that works well in one area can be inadequate with another. The actual solution of a specific problem still requires significant research and development. [2]

Out of the ten sub-categories of digital image processing, listed above, this thesis deals with image enhancement. Here, various enhancement methodologies are used and various inputs are restored using these methods. Image enhancement techniques can be divided into two broad categories:

1. Spatial domain methods, which operate directly on pixels
2. Frequency domain methods, which operate on the Fourier transform of an image.

The rest of the chapter is organized as follows. Document Image Processing and Data Capture are discussed in Section 1.2 and Section 1.3 respectively. The problem definition is described in Section 1.4. Motivation behind carrying out the work is stated in Section 1.5. Organization of the thesis is outlined in Section 1.6.

## 1.2 Document Image Processing

The objective of document image processing is to recognize text and graphics components in images of documents, and to extract the intended information as a human would. Two categories of document image processing can be defined: [3]

- **Textual processing** deals with the text components of a document image. Some tasks here are: determining the skew (any tilt at which the document may have been scanned into the computer), finding columns, paragraphs, text lines, and words, and finally recognizing the text (and possibly its attributes such as size, font etc.) by optical character recognition (OCR).

- **Graphics processing** deals with the non-textual line and symbol components that make up line diagrams, delimiting straight lines between text sections, company logos etc. Pictures are a third major component of documents, but except for recognizing their location on a page, further analysis of these is usually the task of other image processing and machine vision techniques. After application of these text and graphics analysis techniques, the several megabytes of initial data are culled to yield a much more concise semantic description of the document.

Document analysis systems will become increasingly more evident in the form of everyday document systems. For instance, OCR systems will be more widely used to store, search, and excerpt from paper-based documents. Page-layout analysis techniques will recognize a particular form, or page format and allow its duplication. Diagrams will be entered from pictures or by hand, and logically edited. Pen-based computers will translate handwritten entries into electronic

documents. Archives of paper documents in libraries and engineering companies will be electronically converted for more efficient storage and instant delivery to a home or office computer. Though it will be increasingly the case that documents are produced and reside on a computer, the fact that there are very many different systems and protocols, and also the fact that paper is a very comfortable medium for us to deal with, ensures that paper documents will be with us to some degree for many decades to come. The difference will be that they will finally be integrated into our computerized world.

## 1.3 Data Capture

Data in a paper document are usually captured by optical scanning and stored in a file of picture elements, called pixels, that are sampled in a grid pattern throughout the document. These pixels may have values: OFF (0) or ON (1) for binary images, 0–255 for grayscale images, and 3 channels of 0–255 colour values for colour images. At a typical sampling resolution of 120 pixels per centimetre, a $20 \times 30$ cm page would yield an image of $2400 \times 3600$ pixels. When the document is on a different medium such as microfilm, palm leaves, or fabric, photographic methods are often used to capture images. In any case, it is important to understand that the image of the document contains only raw data that must be further analysed to glean the information. [3]

All samples used in this thesis were digitized by scanning through a book scanner, Bookeye 3 (Figure 1.3) consisting of the following features:

- High-resolution CCD image sensors
- Integrated motor-driven book cradle with 100 mm range
- Lamps: LUXEON white LED (light strip: approx. 5000 Lux)
- Resolution: genuine 400 dpi on A1
- Colour depth: 36 bit internal / 24 bit external

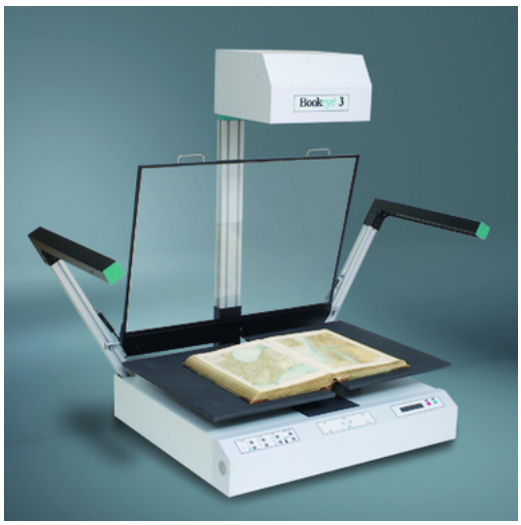


Figure 1.1: The Bookeye 3 Scanner

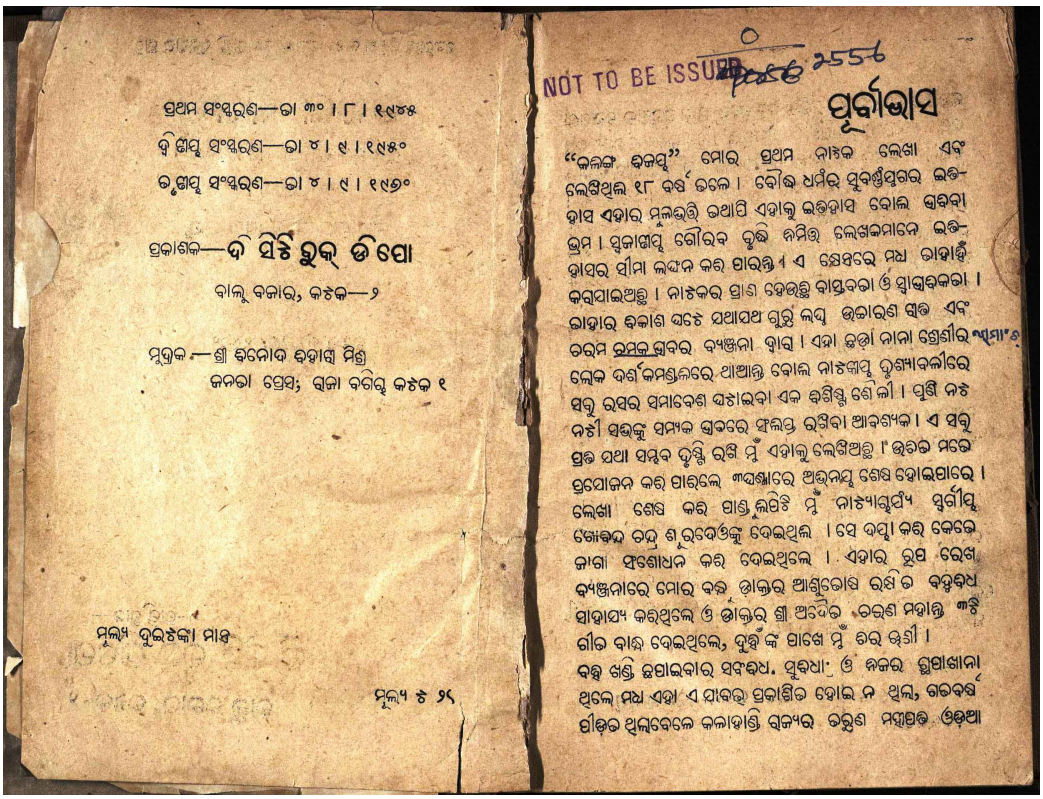
- Interface: 1000 MBit TPC/IP, Scan2Net
- Software: BCS-2
- Format: A1 6350 x 9000 mm

## 1.4 Problem Definition

Document Image Processing has many different methods to accomplish the mentioned tasks. When documents are written on both sides, quite often ink bleeds through the paper. This is a common phenomenon with old documents and low quality paper. With the presence of increased bleed-through, reading and deciphering the text becomes tedious. Bleed-through correction remains one of the vitals part in Document Image Processing.

## 1.5 Motivation

Our work was primarily focused on enhancement of documents available at the central library (Biju Pattnaik Library) of the National Institute of Technology,



ପ୍ରଥମ ସଂସ୍କରଣ—ତା ୩୦ । ୮ । ୧୯୬୫
ଦ୍ୱିତୀୟ ସଂସ୍କରଣ—ତା ୪ । ୯ । ୧୯୫୦
ତୃତୀୟ ସଂସ୍କରଣ—ତା ୪ । ୯ । ୧୯୬୦

ପ୍ରକାଶକ—ଦି ସିଟି ବୁକ୍ ଡିପୋ
ବାଲୁ ବଜାର, କଟକ—୨

ମୁଦ୍ରକ—ଶ୍ରୀ ବନୋଦ ବହାସ୍ୱ ମିଶ୍ର
ଜନତା ପ୍ରେସ୍; ସୁଜା ବଗିଚା କଟକ ୧

ମୂଲ୍ୟ ଦୁଇଟଙ୍କା ମାତ୍ର

ମୂଲ୍ୟ ୫ ୬୯

NOT TO BE ISSUED 2556

## ପୂର୍ବାଭାସ

"କଳଙ୍କ ବଜାର" ମୋର ପ୍ରଥମ ନାଟକ ଲେଖା ଏବଂ ଲେଖିଥିଲ ୧୮ ବର୍ଷ ତଳେ। ବୌଦ୍ଧ ଧର୍ମର ସୁକଶ୍ରୁସ୍ତରର ଇତିହାସ ଏହାର ମୂଳଭିତ୍ତି ତଥାପି ଏହାକୁ ଇତିହାସ ବୋଲ ଜାହିବା ଭ୍ରମ। ସୁଜାଖପୁ ଗୌରବ ବୃଦ୍ଧି ନମିତ୍ତ ଲେଖକମାନେ ଇତିହାସର ସୀମା ଲଙ୍ଘନ କରି ଥାରନ୍ତି। ଏ କ୍ଷେତ୍ରରେ ମଧ୍ୟ ତାହାଙ୍କି କରିଯାଇଅଛୁ। ନାଟକର ପ୍ରାଣ ହେଉଛୁ ବାସ୍ତବତା ଓ ସ୍ୱାଭାବିକତା। ତାହାର ବକାଶ ଘଟେ ଯଥାଯଥ ଗୁରୁ ଲଘୁ ଉଚ୍ଚାରଣ ସ୍ୱର ଏବଂ ଚରମ ରସଜ୍ଞ ସ୍ୱର ବ୍ୟଞ୍ଜନା ଦ୍ୱାରା। ଏହା ଛଡ଼ା ନାନା ଶ୍ରେଣୀର ଲୋକ ବର୍ଗକମଣ୍ଡଳରେ ଥାଆନ୍ତି ବୋଲ ନାଟ୍ୟାପୁ ଦୃଶ୍ୟାବଳୀରେ ସବୁ ରସର ସମାବେଶ ଘଟାଇବା ଏକ ବିଶିଷ୍ଟ ଶୈଳୀ। ପୂର୍ଣ୍ଣ ନଟ ନଟୀ ସଭଙ୍କୁ ସମ୍ୟକ ଜ୍ଞାତରେ ଫୁଲପ୍ତ ରଖିବା ଆବଶ୍ୟକ। ଏ ସବୁ ପ୍ରତି ଯଥା ସମ୍ଭବ ଦୃଷ୍ଟି ରଖି ମୁଁ ଏହାକୁ ଲେଖିଅଛୁ। ଜଣେ ମଞ୍ଚେ ପ୍ରଯୋଜନ କରି ପାରିଲେ ଶବ୍ଦାଗାରେ ଅଭିନୟ ଶେଷ ହୋଇପାରେ। ଲେଖା ଶେଷ କରି ପାଣ୍ଡୁଲିପିଟି ମୁଁ ନାଟ୍ୟାଗୁର୍ଯ୍ୟ ସ୍ୱର୍ଗୀୟ ଗୋବିନ୍ଦ ଚନ୍ଦ୍ର ଶ୍ରୀଗୁରୁଙ୍କୁ ଦେଇଥିଲ। ସେ ଦୟା କରି କେତେ ଜାଗା ସଂଶୋଧନ କରି ଦେଇଥିଲେ। ଏହାର ରୂପ ରେଖା ବ୍ୟଞ୍ଜନାରେ ମୋର ବନ୍ଧୁ ଡାକ୍ତର ଆଶୁତୋଷ ରକ୍ଷିତ ବହୁବିଧ ସାହାଯ୍ୟ କରିଥିଲେ ଓ ଡାକ୍ତର ଶ୍ରୀ ଅରବିନ୍ଦ ଚରଣ ମହାନ୍ତି ମଧ୍ୟ ଗୀତ ବାନ୍ଧ ଦେଇଥିଲେ, ଦୁହିଁଙ୍କ ପାଖେ ମୁଁ ଚିର ଋଣୀ।
ବହୁ ଖଣ୍ଡି ଛପାଇବାର ସଂବନ୍ଧ. ସୁବଧା: ଓ ଜଜର ଘୁଣାଖାନା ଥିଲେ ମଧ୍ୟ ଏହା ଏ ଯାବତ୍ ପ୍ରକାଶିତ ହୋଇ ନ ଥିଲ, ଗତବର୍ଷ ଗାନ୍ଧିତ ଥିଲବେଳେ କଳାଡାଣ୍ଡି ବଜ୍ୟର ଚତୁର ମସ୍ତାନ୍ତ ଓଡ଼ିଆ

Figure 1.2: Sample of data captured with the Bookeye 3 scanner under Project OaOb

Rourkela, Orissa. A variety of old manuscripts, documents and books acquired through the Open Access to Oriya Books (Project OaOb) was found to pose four problems common to all documents through scanning as seen in Figure 1.3

1. *Introduction of skew*: During scanning, skew is inevitably introduced into the process due to lack of precision in placement of the document or the book on the scanner cradle.

2. *Marginal and dimensional discrepancies*: Scanned documents, mostly handwritten were found to have persistent differences in the widths of margins on all four boundaries of a page. Moreover, these documents were also inconsistent in page dimensions.

3. *Bleed-through*: Most old documents with poor quality or degraded paper suffered seepage of ink from one side of the page to the other.

4. *Marks and artefacts*: Ink blotches, human made markings and loss of data due to physical damage of the document were other problems faced.

Software used for enhancement viz. BCS 2, ABBYY Finder and Adobe Acrobat where found to deal with all of the above stated problems except bleed-though to a considerable level. Bleed-through distortion required the specialist to work on files individually to enable their removal which still did not result in desired results. It was imperative to consider a stronger, quicker, more accurate batch processing technique to the problem and that has hence been the motivation for this project.

## 1.6 Thesis Organization

The rest of the thesis is organized as follows:

> **Chapter 2** introduces thresholding and describes how optimal and adaptive thresholding can be applied to remove bleed-through to a considerable level.

**Chapter 3** outlines the structure, organization and working of the bleed-through removal algorithms described within. A widely used technique for removal which works on registration using optimization is explored. In addition, a method for efficient registration of documents through cross-correlation is proposed.

**Chapter 4** presents the concluding remarks, with scope for further research work.

# Chapter 2

# Thresholding

## 2.1 Introduction

Thresholding is a simple form of image segmentation through which a binary image can be obtained from a grayscale image. During the process, individual pixels in an image are categorized as "object" pixels or "background" pixels if their value is greater than some threshold value or otherwise . This convention is known as *threshold above*. Other conventions include *threshold below*, which is opposite of threshold above; *threshold inside*, where a pixel is labelled "object" if it has a value between two thresholds and its opposite as *threshold outside*. [4] Generally, an object pixel is given a value of 1 while a background pixel is given a value of 0. Finally, a binary image is created by colouring each pixel white or black, depending on the pixel's label. (Figure 2.1)

The key factor behind the thresholding process to choose the threshold value. Several methods for choosing a threshold are available; one can manually choose a threshold value, or a thresholding algorithm can be used to compute the same automatically, *known as automatic thresholding*. [4] Choosing the mean or median value would be a fairly simple method, the idea being that if the object pixels are brighter than the background, they should also be brighter than the average. This works well in a noiseless image with uniform background and object values.

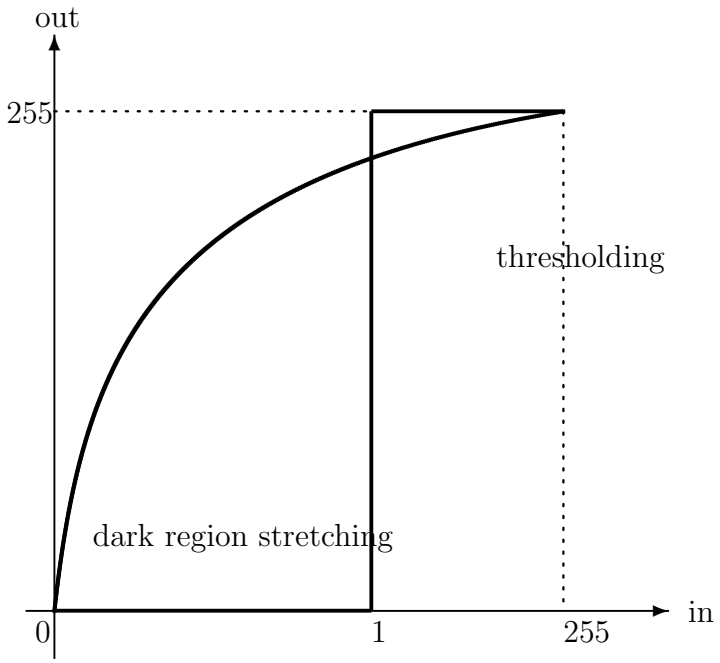



Figure 2.1: Graph depicting thresholding

A more accurate approach would be to create a histogram of the image pixel intensities and use the valley point as the threshold. The approach assumes that there is some average value for the background and object pixels, but that the actual pixel values have some variation around these average values. But this may be computationally intensive, and image histograms may not produce clearly defined valley points, often making the selection of an accurate threshold difficult. One method that does not require much specific knowledge of the image, and works well against image noise is the following iterative method:

1. An initial threshold ($T$) is chosen, either randomly or according to any other method desired.

2. The image is segmented into object and background pixels by creating two sets:

$$G_1 = \{\ f(m,n) : f(m,n) > T\ \}$$

$$G_2 = \{\ f(m,n) : f(m,n) \leq T\ \}$$

where *f(m,n)* is the value of the pixel located in the $m^{th}$ column, $n^{th}$ row

3. The average of each set is then computed.
   (a) $m_1$ = average value of $G_1$
   (b) $m_2$ = average value of $G_2$
4. A new threshold is obtained with the average of $m_1$ and $m_2$
   $T' = (m_1 + m_2)/2$
5. Go to step two and now using the new threshold repeat until the new threshold matches the previous one i.e. until convergence has been reached.

This iterative algorithm is a special one-dimensional case of the k-means clustering algorithm, which has been proven to converge at a local minimum. Hence, a different initial threshold may give a different final result.

## 2.2 Otsu's Method

Otsus algorithm makes use of the zeroth and first order cumulative moments of the grey level histograms to predict a threshold value. For an image A, with $k$ grey levels, we can predict a threshold value, $k_{thresh}$ which divides the whole image into two classes of pixels. Let the mean and the variance of the object and background with respect to any arbitrary threshold value $t$ be denoted by ($m_1$, $v_1$) and ($m_2$, $v_2$) respectively, and $p$ be the cumulative probability of the foreground. Then, Otsus algorithm proposes an optimum threshold grey level $k_{thresh}$ such that [8]

$$\alpha(k_{thresh}) = max(\alpha(t)) \tag{2.1}$$

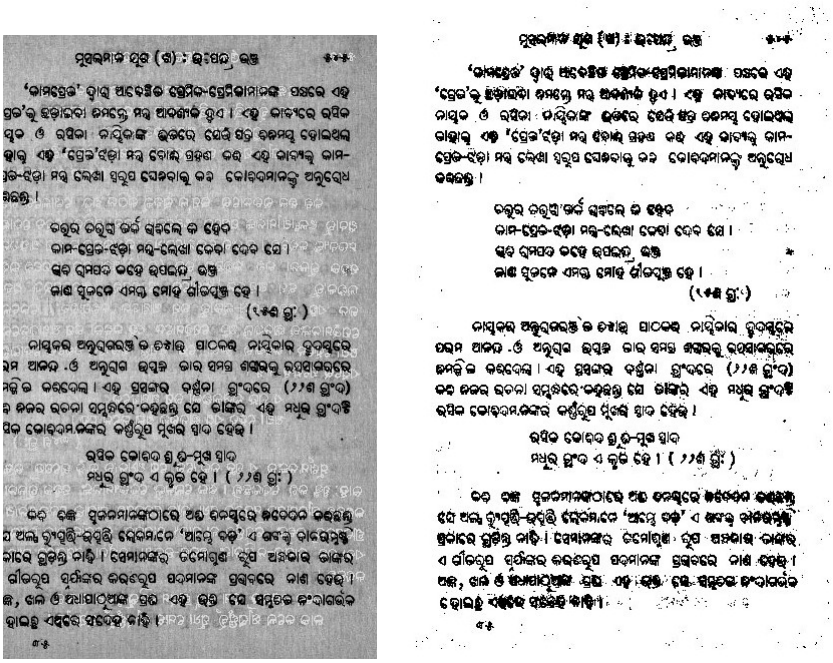
$$\alpha(t) = \frac{p(1-p)(m_1 - m_2)^2}{pv_1 + (1-p)v_2} \tag{2.2}$$

## 2.3 Bleed-through Removal through Thresholding

The document suffering from bleed through can be subjected to basic thresholding and Otsus method for optimal thresholding to enhance readability. However these techniques fail to remove bleed through from low contrast document images. Moreover, it is also found to result in loss of data in many cases.
An alternative would be to apply **adaptive thresholding** wherein a different threshold may be used for different regions in the image. This is also known as *local or dynamic thresholding.*

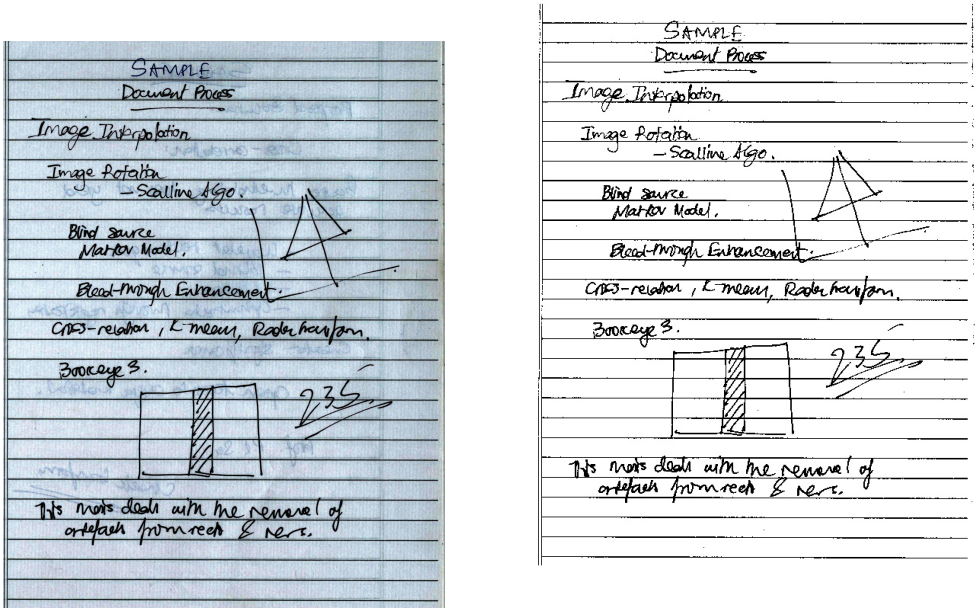
The process of removal of bleed-through in a scanned manuscript is shown below. Figure 2.2(a) shows the original scanned document which has bleed-though distortion. Figure 2.2(b) shows the outcome of Otsu's optimal thresholding; a good level of bleed-through removal coupled with loss in data which, is not desirable.

On the contrary, thresholding produces near accurate results in a different document. (Figure 2.3)



(a) Original scanned image (b) Thresholding to remove bleed-through

Figure 2.2: Sample of bleed-through removal by thresholding in a scanned Project OaOb document




(a) Original scanned image (b) Thresholding to remove bleed-through

Figure 2.3: Sample of bleed-through removal by thresholding in text from a book with high quality paper

# Chapter 3

# Bleed-Through Removal Algorithm

## 3.1 Overview

The use of a simple intensity threshold to eliminate the bleed does not yield desirable results. The intensity of the bleed through is often very similar to the intensity of the original writing such that they are too similar to dissociate them with a threshold. Hence, algorithms to tackle the same have to be used. Such algorithms are divided in two parts: registration and restoration. The registration involves digitizing both the recto and the document and attempting to perfectly make the recto and the flipped verso overlap. By having the bleed through perfectly aligned with the original writing, it will be possible to eliminate the bleed through and restore the document using thresholds.

## 3.2 Registration

Image registration is the process of transforming the different sets of data into one coordinate system. Registration is necessary in order to be able to compare or integrate the images that have been obtained in different measurements while scanning. The registration algorithms for bleed-through may fall under either of:

- **Intensity-based or feature-based**: Image alignment algorithms are classified into intensity-based and feature-based. One of the images is referred to as the reference and the other as the target or sensed. Registration involves spatially transforming the target image to align with the reference image. While intensity-based methods compare intensity patterns in images via correlation metrics, feature-based methods find correspondence between image features such as points, lines, and contours. Intensity-based methods register entire images or subimages whose centers are treated as corresponding feature points. Feature-based methods gives a correspondence between a number of points in images. A transformation is then determined to map the target image to the reference images knowing the correspondence between a number of points in images, thereby establishing point-by-point correspondence between the reference and target images.

- **Transformation models**: They maybe used to relate the target image space to the reference image space. They include linear transformations, which consist of rotation, scaling, and other affine transforms. Linear transformations are global in nature and thus cannot model local geometric differences between images.The second category of transformations allow 'elastic' or 'nonrigid' transformations. They are capable of locally warping the target image to align with the reference image. Nonrigid transformations include radial basis functions (thin-plate or surface splines, multiquadrics, and compactly-supported transformations), physical continuum models (viscous fluids), and large deformation models (diffeomorphisms).

- **Spatial and frequency domain methods**: Spatial methods operate in the image domain, matching intensity patterns or features in images. Several of the feature matching algorithms are extensions of traditional techniques for manual image registration, wherein an operator chooses corresponding control points (CPs) in images. When the number of control points exceeds the minimum required to define the appropriate transformation model, iterative algorithms can be used to robustly estimate the parameters of a

particular transformation type (e.g. affine) for registration of the images. Frequency-domain methods find the transformation parameters while working in the transform domain. They work for simple transformations, such as translation, rotation, and scaling. Applying the Phase correlation method to a pair of images produces a third image which contains a single peak which corresponds to the relative translation between the images. Unlike many spatial-domain algorithms, the phase correlation method is resilient to noise. In addition phase correlation uses the fast Fourier transform to compute the cross-correlation between the two images, generally resulting in better performance. The method can be extended to determine rotation and scaling differences between two images by first converting the images to log-polar coordinates. Due to properties of the Fourier transform, the rotation and scaling parameters can be determined in a manner invariant to translation.

### 3.2.1 Registration through Optimization

#### Introduction

This chapter explores a popular and widely used method for the registration of documents as described in Section 3.2. First, both sides of the document are digitized. The verso is then flipped and corrected so that the bleed through coordinates of one side correspond exactly with the coordinates of the original writing on the other. This is done because both sides to handle any difference in shift or rotation between the two. An affine transformation of six parameters is used for the same whose parameters are found by optimizing the alignment process.

#### Algorithm

Let us denote the recto of the document as $f_r(x, y)$ and the flipped verso as $f_v(x, y)$. To obtain an ideal flipped verso, $f_v^I(x, y)$, some transformation has to be applied to $f_v(x, y)$. [5] This is done using an affine transformation, $A_t$

$$f_v^I(x,y) = f_v(x,y) * A_t \tag{3.1}$$

This affine transformation consists of six parameters: $t_{11}, t_{12}, t_{21}, t_{22}$ are used to correct the rotation while $t_{13}$ and $t_{23}$ are used to correct the horizontal and vertical shifts respectively. To find those parameters that will correct the verso perfectly, we use the *difference*. The difference is viewed as the difference that remains between the recto and the flipped transformed verso when they are aligned. It is a scalar and is denoted by:

$$difference = ||f_r - A_t * f_v||^2 \tag{3.2}$$

It is evident that the smaller the difference the more accurate the alignment will be. This depends on the choice of parameters for the transformation. An estimation of the parameters can be found by solving the following optimization problem:

$$t_e = \arg_t \min ||f_r - A_t * f_v||^2 \tag{3.3}$$

This equation shows that the affine transformation parameters have to minimize the difference. It returns $t_e$, the estimated best parameters and using those we obtain our estimated ideal verso:

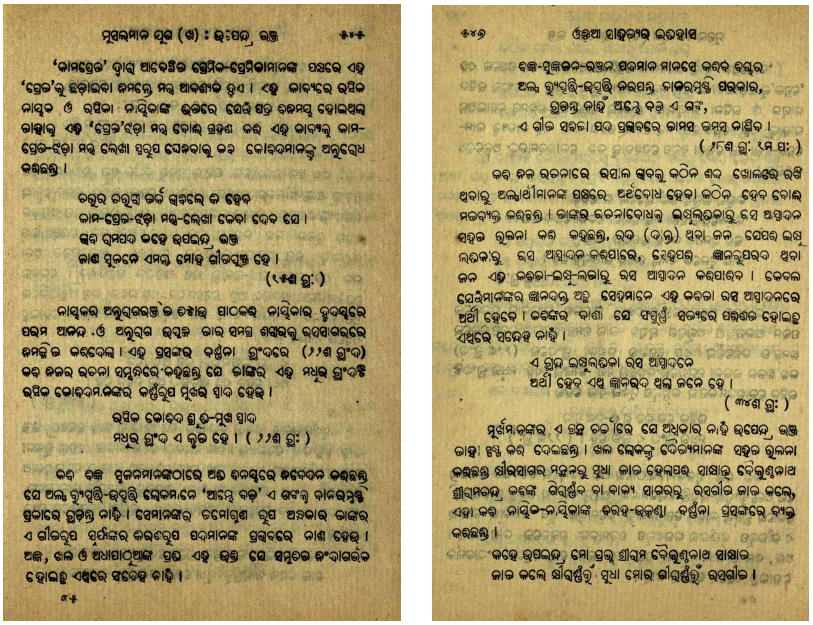
$$f_v^I = At_e * f_v \tag{3.4}$$

The detailed process of bleed-through removal by registration by optimization followed by restoration is depicted by figures 3.1(a) and 3.1(b) which are the scanned recto and verso images. 3.1(c) shows the registered and restored image.

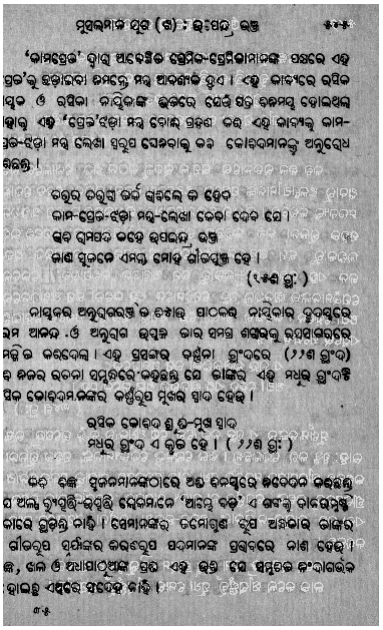
### 3.2.2 Registration using Cross-correlation

#### Overview

Objects can be classified in ways other than measuring the dimensions or shapes of objects. Image matching techniques compare portions of images against one



(a) Scanned recto image (b) Scanned verso image



(c) Registered and restored recto

Figure 3.1: Bleed-through removal by registration by optimization in a Project OaOb document

another. This is done by a process known as spatial (or two dimensional) **cross correlation**. Referred to as matched filtering, this technique involves a pixel-by-pixel comparison of a small reference image containing an object of interest with an image under analysis. The result obtained shows image showing bright spots where the image matches the reference image. The brighter the spot, the better the match. Hence, when the brightness is great enough, a match in objects is determined. Quite often, this is obtained as a graph showing a peak where the brightness is observed to be maximum. [6]

### Registration Process

The matched filter is implemented in a similar manner to pixel group processes but instead of a mask of weight values, matched filters use an image mask composed of a reference image. The image mask is a small image depicting the object that we wish to find in an image. The mask dimensions are usually greater than the 3 X 3 and 5 X 5 pixel sizes used in spatial convolution. Rather, they can be any size; the only limitation is the computational effort necessary to compute each output pixel value.

The mechanics are similar to spatial convolution in that the image mask is moved over the input image, pixel by pixel, placing resulting pixels in the output image. At each pixel location, the input pixels are compared with the pixels of the image mask and a resulting output pixel is created, the cross correlation coefficient, where a dark value represents a poor match between the two and a bright output value indicates a good match. Similarly, for a graphical plot a high peak represents a good match and vice versa. The resulting output pixel value is computed as the sum of the input pixels in the group, each multiplied by its respective image mask pixel value.

The resulting output values can may become enormous when all brightness values tend to be equal. However, the output values will be relatively small when there is not good correlation between the two.

Basic cross correlation can be give as follows: [6]

$$O(x,y) = \sum_i \sum_j I(x+i, y+j) * T(i,j) \tag{3.5}$$

Where $O(x,y)$ is the cross correlation function, $I(x,y)$ is the image function and $T(i,j)$ is the template. Usually the template is a sub-image smaller than $I(x,y)$ and the summation is over that area of $I(x,y)$ overlapped by $T(i,j)$. The maximum value of $O(x,y)$ is where the image and the template best match and the maximum over all categories gives us the classification.

**Steps**

1. *Read the vecto and verso images*

2. *Choose subregions in the verso image*: It is important to choose regions that are similar. Figure 3.1(b) is the template here.

3. *Perform normalized cross-correlation and find the coordinates of the max peak*: Calculate the normalized cross-correlation and display it as a surface plot. The peak of the cross-correlation matrix occurs where the sub images are best correlated. (Figure 3.2(a))

4. *Find the total offest between the images*: The total offset or translation between images depends on the location of the peak in the cross-correlation matrix, and on the size and position of the sub images.

5. *Align the verso on the recto image*: Pad figure 3.1(b) to overlay on figure 3.1(a), using the offset determined above.

6. *Verify dimensions*: If both the images are not of the same size, resize accordingly.

7. *Perform restoration*: As detailed in Section 3.3

Registration by cross-correlation of figures 3.1(a) and 3.1(b) is shown in figure 3.2(a) which a plot of the peak which determines the offset. Cross-correlation

followed by an intensity mapping is then performed as shown in figure 3.2(b). The image is finally restored (Figure 3.2(c)).

## 3.3 Restoration

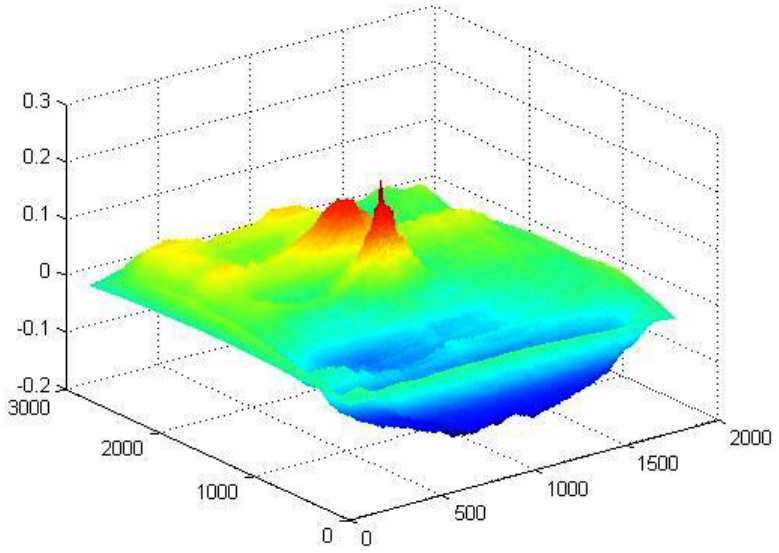
After the registration process is complete, the document can undergo restoration. The restoration process for a document for a bleed through removal algorithm aims at replacing the pixels of the recto corresponding to bleed through with the background. The pixels of the recto can be classified under one of four categories: [5]

1. There is recto writing but no bleed through.
2. Bleed through exists but recto writing does not.
3. There is neither recto writing nor bleed through
4. Both recto writing and bleed through exist.

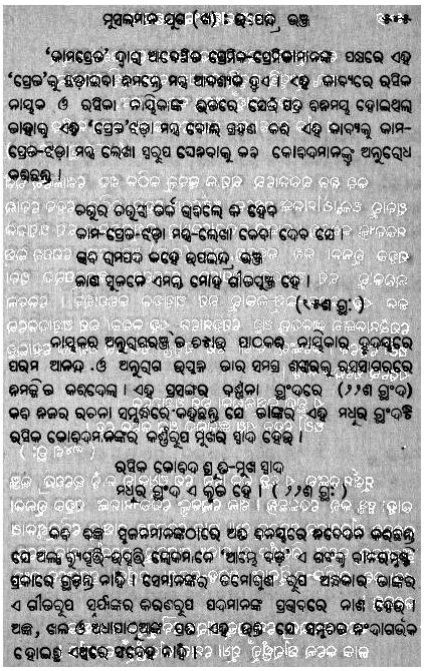
Using thresholds, we can identify when a pixel falls in category 2 and we can then replace it with the background value. This explains why it is necessary to obtain a perfect alignment in the registration process. The bleed through pixels don't correspond to the pixels of the recto writing if we try to find in which category the pixel fall, with a bad alignment. They have to correspond exactly like in the original document, as it was before scan.

## 3.4 A complete comparative example

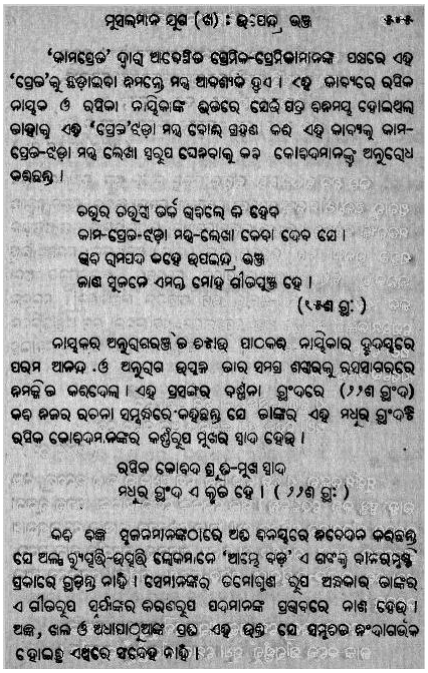
The following images give a complete comparative analysis of the three methods implemented in the above sections.




(a) Surface plot

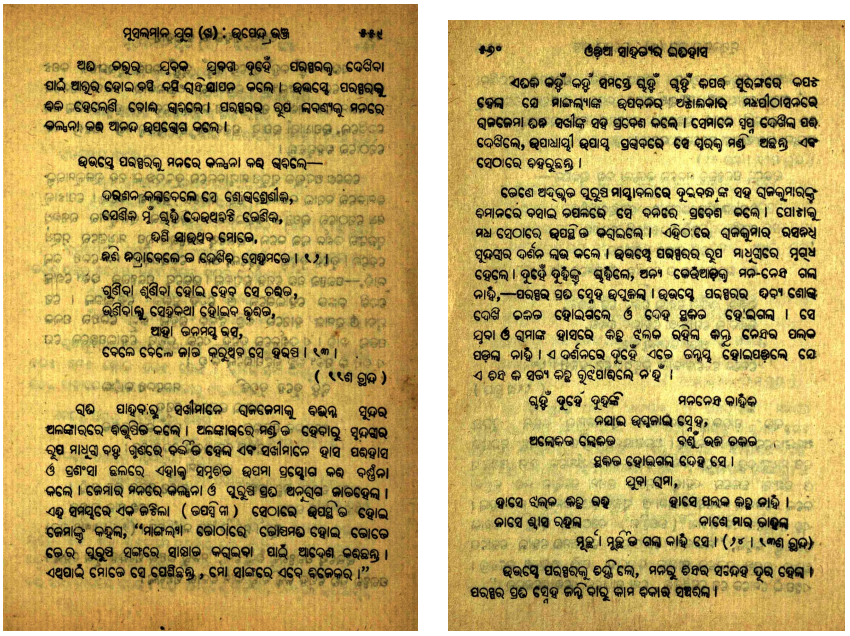


(b) Correlation and intenstiy mapping



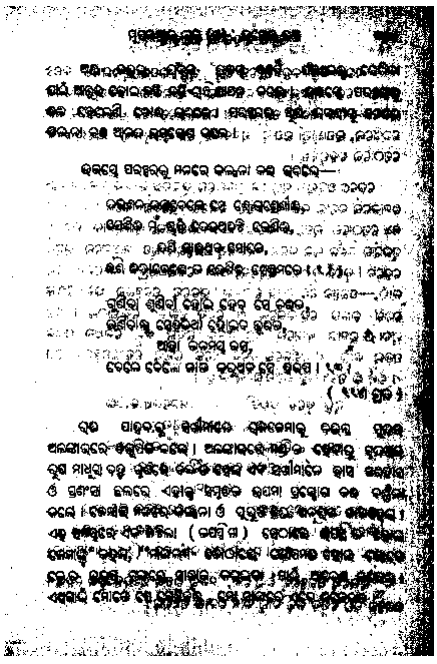
(c) Restored recto

Figure 3.2: Bleed-through removal by registration using correlation in a Project OaOb document



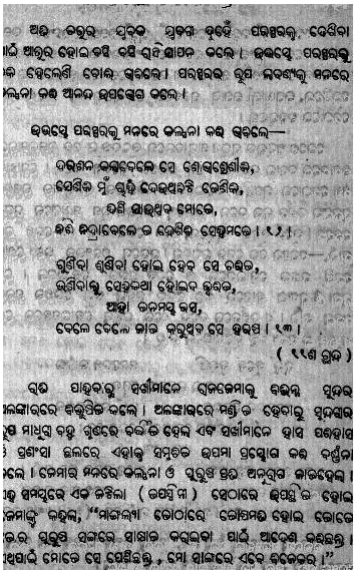
(a) Scanned recto image (b) Scanned verso image

Figure 3.3: Scanned recto and verso images of an Oriya manuscript



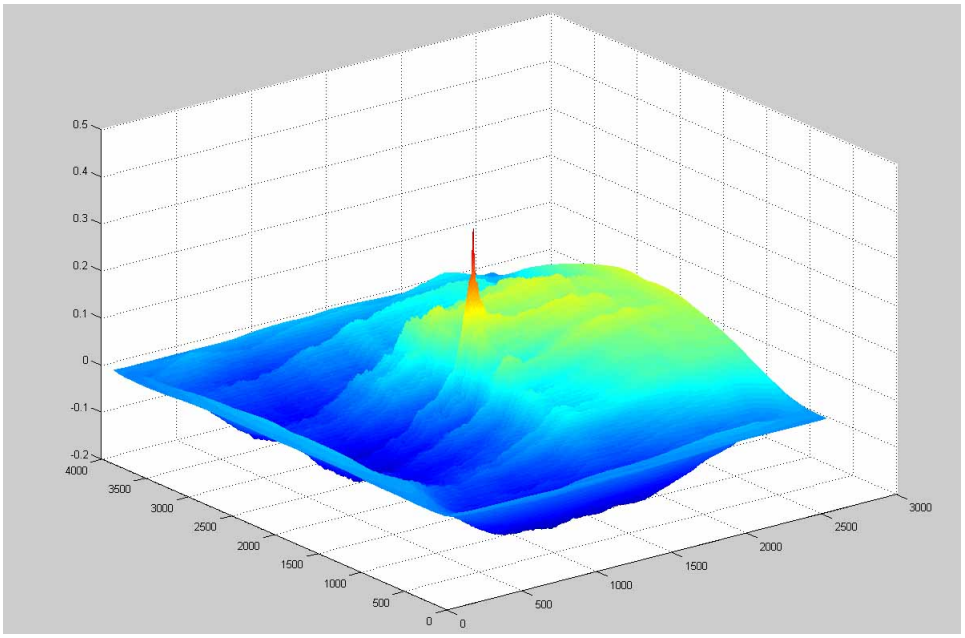
(a) Thresholding to remove bleed-through

Figure 3.4: Bleed-through removal by thresholding

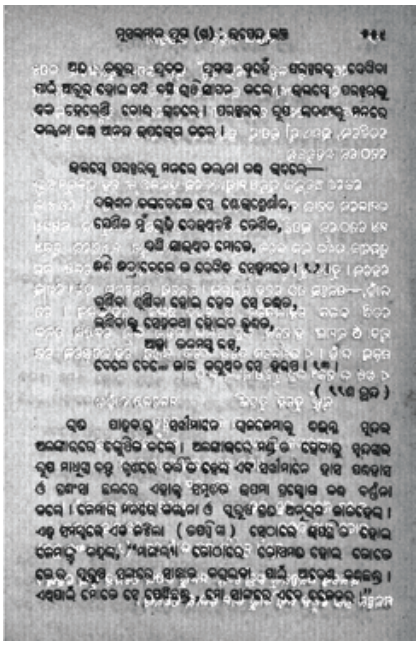


(a) Registered and restored recto

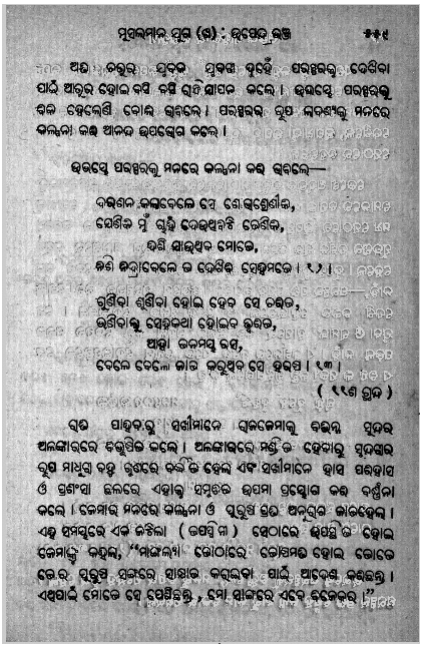
Figure 3.5: Bleed-through removal by registration by optimization




(a) Surface plot



(b) Correlation and intenstiy mapping



(c) Restored recto

Figure 3.6: Bleed-through removal by registration using correlation

# Chapter 4

# Conclusions

The work in this thesis, primarily focuses on bleed-through removal in scanned images. The work reported in this thesis is summarized in this chapter. Section 4.1 lists the achievements of the work. 4.2 lists the limations and 4.3 provides some scope for further development.

## 4.1 Achievements

Two known methods for bleed-through removal from scanned images viz. thresholding and registration by optimization were studied under great detail and implemented. In addition, a third new method involving registration by correlation was proposed and implemented on the same samples. The proposed method was found to yield desirable results at par with the existing techniques that were explored and may be considered as an alternate approach for batch processing documents suffering from bleed-through but not having rotational skew. Above all, this was found to tackle the bleed-through problems of the scanned manuscripts of Project OaOb at the Biju Pattnaik Library of the National Institute of Technology, Rourkela, Orissa.

## 4.2 Limitations of the work

The following limitations were encountered during the course of this project:

Although a very simple and efficient method for bleed-through removal, thresholding has a number of drawbacks. It fails to remove bleed-through from low contrast document image. When Otsu's algorithm is used for finding the optimal threshold value in batch processing, it results in loss of detail in the images. Manual effort is need to zero in on the most appropriate threshold intensity which will conserve detail.

The registration by optimization method is very efficient in terms of time complexity for alignment of recto and verso images but accuracy in alignment was not achieved for large image sizes.

Registration by cross-correlation was effective in alignment of recto and verso images but the implementation of the technique was heavy in terms of memory requirements. Also this method fails when there is rotational skew between the recto and verso images.

## 4.3 Further Development

A method for skew detection and correction can be combined with the proposed method for a complete solution for removal of bleed through. Also this method can be extended to incorporate colour information, for the restoration and storage of scanned manuscripts. A batch processing technique incorporating the procedure can also be developed for automated removal of bleed through from a the set of scanned images. In addition, solutions to the problems common to all documents digitized through scanning as listed in Section 1.5, can be developed.

# Bibliography

[1] R. C. Gonzalez and R. E. Woods. *Digital Image Processing.* Addison Wesley, 1992.

[2] Anil K. Jain. *Fundamentals of Digital Image Processing.* Prentice Hall of India, 2008.

[3] Rangachar Kasturi, Lawrence O'Gorman, and Venu Govindaraju. Document image analysis: A primer. *Sadhana Vol. 27, Part 1*, February 2002.

[4] L. G. Shapiro and G. C. Stockman. *Computer Vision.* Prentice Hall, 2002.

[5] Yohan Bienvenue. Registration of two-sided documents suffering from bleed through. Technical report, 2001.

[6] H. Stahlberg, R. Zumbrunn, and A. Engel. *Digital Image Processing in Natural Sciences and Medicine.* 2002.

[7] Eric Dubois and Anita Pathak. Reduction of bleed-through in scanned manuscript documents.

[8] Sahil Mahaldar and Serene Banerjee. Enhanced bleed through removal using normalized picture information based measures. Hewlett-Packard Development Company, L.P., July 2009.